DOI 10.1002/pmic.201300118 *Proteomics* 2013, *13*, 1548–1550

Commentary

Resilience in the proteomics data ecosystem: How the field cares for its data

Lennart Martens[1,2]

[1] Department of Medical Protein Research, VIB, Ghent, Belgium
[2] Department of Biochemistry, Ghent University, Ghent, Belgium

The public dissemination of data is an integral part of the life sciences. In the field of proteomics too, data sharing has taken off over the last few years, with the first downstream uses of these data quickly gaining prominence. At the same time, the recent unfortunate demise of two repositories, NCBI Peptidome and ProteomeCommons Tranche, has shown the frailty of such data gathering efforts. Heroic efforts by the PRIDE and Peptidome teams to rescue the Peptidome data have now ensured their continued availability to the field, and alternatives have already been put in place for Tranche. But with public data increasingly at the hub of the life sciences, it is a good time to look at the proteomics data ecosystem in some more detail.

Received: March 22, 2013
Accepted: March 26, 2013

Keywords:
Bioinformatics / Public repositories / Standards

Ever since the first 3D protein structures were made freely available to the world, public data dissemination has been a defining aspect of the modern life sciences. Some 30 years later, open access to the human genome once and for all etched this unconditional sharing of information into the foundations of the field. The impact of this policy has been nothing short of impressive, including giving birth to entirely new fields such as proteomics. Unsurprisingly, data sharing has in turn also gained increasing traction in the field of proteomics, with the first tangible outcomes of data reuse quickly gaining prominence. Indeed, although the inherent heterogeneity of the data always needs to be taken into account when performing such analyses [1,2], downstream processing of publicly available proteomics data has already resulted in new knowledge [3,4] and even in new types of resources [5,6]. Data sharing in such a data-intensive field is not a trivial undertaking however, since it requires substantial investment in infrastructure. Several databases have correspondingly been built for MS-based proteomics data, with GPMDB [7], PeptideAtlas [8], and PRIDE [9] among the first, later followed by Tranche and Peptidome [10]. A distinction can be made between true repositories on the one hand (including PRIDE, Tranche, and Peptidome) that store data exactly as submitted, and resources (PeptideAtlas, GPMDB) that are built on reprocessed data using an in-house pipeline. For journals, the most interesting systems are typically the repositories, as they can maintain the data supporting a publication in their original form, while many downstream users consider the homogeneity and filtering applied by the reprocessing resources a strong benefit. With all these systems in place, one would expect proteomics to be in good shape toward sustaining a healthy data ecosystem. Yet in a relatively short time span, two repositories were discontinued, first Peptidome in 2011 and then Tranche a few months ago. The untimely demise of these systems dramatically illustrated the inherent frailty of any effort to host and serve data for a long period of time, and the fact that Peptidome was an NCBI resource made it clear that even large organizations with data collection and dissemination at their core are not immune from resource collapse. Indeed, examples such as these illustrate the importance of guarding public data against a single point of failure;

Correspondence: Professor Lennart Martens, Department of Medical Protein Research, Universiteit Gent – VIB, A. Baertsoenkaai 3, B-9000 Gent, Belgium
E-mail: lennart.martens@vib-ugent.be
Fax: +32 9 264 94 84

© 2013 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

anyone who has ever lost a large digital photo collection or an important set of documents due to the failure or theft of a single machine can attest to the importance of data redundancy. To this end, the major stakeholders in proteomics data sharing have created the ProteomeXchange consortium (http://www.proteomeXchange.org) to provide a single point for data submission, but multiple points of data storage and dissemination. In fact, anyone interested can register to receive structured e-mails, RSS feeds, or even Twitter updates from ProteomeXchange whenever a new data set is made available.

However, at this point one could wonder aloud whether all these data should in fact be kept in databases at all. The argument that is typically put forward is that instruments constantly get better, and that yesterday's data are therefore very much yesterday's news. The conclusion then is that only certain data sets, for instance those derived from highly precious or unique samples, are worth storing long term. Yet there is a flaw in this argument, and that is that it overlooks the immense benefit of having large amounts of mostly independent data available for downstream analysis. Indeed, the examples of data reuse cited above show that orthogonal interrogation of public data sets can turn up unexpected and novel findings [3, 4], while resources such as MOPED [5] and PaxDB [6] expressly rely on large, multiexperiment data sets to create a comprehensive, quantitative view on a whole proteome. It is of particular note that these latter resources do not uniquely target the proteomics community as users, but actively reach out to the much broader community of biological and biomedical researchers. It is highly likely that the real impact of public proteomics data will be found there, in the influence that this amassed information has on downstream fields of research. In that respect, proteomics itself is a case in point, existing solely by grace of the public availability of extensive protein sequence databases derived largely from freely available sequenced genomes. Like in all life sciences therefore, a healthy data ecosystem in proteomics will hold many future benefits, few of which can be envisioned at the start.

In order to maintain the health of the proteomics data ecosystem then, and as reported by Csordas et al. in this issue, the PRIDE and Peptidome teams joined forces to transfer all data from Peptidome to PRIDE, a herculean but largely transparent effort that is worthy of the praise of the community. As a result, the discontinuation of Peptidome has had no significant effect on the availability of its data holdings. It is however interesting to read in the report by Csordas et al. [11] about the various issues encountered during data transfer. One quickly realizes that the field still needs to invest in the pervasive standardization of data (e.g., mzML [12], mzIdentML [13]) and in the reporting of metadata according to minimal reporting requirements [14–16]. Meanwhile, for raw data storage [17], the role of Tranche as the repository of choice has been taken over by EMBL-EBI [18], with a second effort led by Nuno Bandeira at UCSD, dubbed Mass Spectrometry Interactive Virtual Environment (MassIVE), now also well underway. Similar to the regenerative capacity of the mythical hydra, the removal of one important resource in the proteomics data ecosystem has resulted in the emergence of two new resources to replace it, displaying a resilience that inspires confidence for the future of data sharing in proteomics.

Indeed, while the closures of Tranche and Peptidome have come as two consecutive blows to the health of the data ecosystem in the field of proteomics, the field has rebounded quickly and decisively, showing impressive vigor in preserving structured data from oblivion, while energetically creating new resources to replace discontinued ones. In this first and challenging test, the field has thus shown that it can be trusted to handle the data that it generates responsibly, and that it is resolute in its ambition to provide uninterrupted public access to these data. Given the importance of publicly available data in the history of the field, and based on the first glimpses of what may yet lie in its future, I firmly believe that this is a very, very good thing!

The author acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), and the PRIME-XS and ProteomeXchange projects funded by the European Union 7th Framework Program under grant agreement numbers 262067 and 260558, respectively.

The author has declared no conflict of interest.

References

[1] Gonnelli, G., Hulstaert, N., Degroeve, S., Martens, L., Towards a human proteomics atlas. *Anal. Bioanal. Chem.* 2012, *404*, 1069–1077.

[2] Foster, J. M., Degroeve, S., Gatto, L., Visser, M. et al., A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* 2011, *11*, 2182–2194.

[3] Hahne, H., Moghaddas Gholami, A., Kuster, B., Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. *Mol. Cell. Proteomics* 2012, *11*, 843–850.

[4] Matic, I., Ahel, I., Hay, R. T., Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites. *Nat. Methods* 2012, *9*, 771–772.

[5] Kolker, E., Higdon, R., Haynes, W., Welch, D. et al., MOPED: Model Organism Protein Expression Database. *Nucleic Acids Res.* 2012, *40*, D1093–D1099.

[6] Wang, M., Weiss, M., Simonovic, M., Haertinger, G. et al., PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* 2012, *11*, 492–500.

[7] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2004, *3*, 1234–1242.

© 2013 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[8] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. et al., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* 2005, *6*, R9.

[9] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537–3545.

[10] Slotta, D. J., Barrett, T., Edgar, R., NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. *Nat. Biotechnol.* 2009, *27*, 600–601.

[11] Csordas, A., Wang, R., Ríos, D., Reisinger, F. et al., From Peptidome to PRIDE: public proteomics data migration at a large scale. *Proteomics* 2013, *13*, 1692–1695.

[12] Martens, L., Chambers, M., Sturm, M., Kessner, D. et al., mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics* 2011, *10*, R110.000133.

[13] Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O. et al., The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* 2012, *11*, M111.014381.

[14] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A. et al., The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 2007, *25*, 887–893.

[15] Montecchi-Palazzi, L., Kerrien, S., Reisinger, F., Aranda, B. et al., The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics* 2009, *9*, 5112–5119.

[16] Medina-Aunon, J. A., Martinez-Bartolome, S., Lopez-Garcia, M. A., Salazar, E. et al., The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards. *Mol. Cell. Proteomics* 2011, M111.008334.

[17] Martens, L., Nesvizhskii, A. I., Hermjakob, H., Adamski, M. et al., Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* 2005, *5*, 3501–3505.

[18] Editorial. A home for raw proteomics data. *Nat. Methods* 2012, *9*, 419.

© 2013 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim